Are LLMs Reliable Judges? Evidence from Subjective, High-Stakes Assessment

Caiyun Wang[1], Yang You[2], Yuxiao Sun[3], Chenyu Gu[4], Kai Chen[5], Zhen Liu[5], Zequn Liu[5*], and Yuanyi Zhen[5*]

[1] Tsinghua University, Zhongguancun Academy
[2] Harbin Institute of Technology, Zhongguancun Academy
[3] Beijing Normal University, Zhongguancun Academy
[4] Beihang University, Zhongguancun Academy
[5] Zhongguancun Academy, Zhongguancun Institute of Artificial Intelligence
[*]Corresponding authors: liuzequn@bza.edu.cn, zhenyuanyi@bza.edu.cn

Abstract. Large language models (LLMs) have been increasingly adopted as automated evaluators, giving rise to the LLM-as-a-judge paradigm. However, whether they can reliably serve as judges in subjective, multi-dimensional, and high-stakes settings remains an open question. In this paper, we construct a novel, real-world dataset of doctoral candidate assessments for AI leadership potential. This expert-curated setting, characterized by structural disagreement among human raters, provides a unique natural testbed to probe the behavior of automated evaluators. Leveraging this dataset, we evaluate multiple LLMs under varying conditions and extend our analysis to multi-agent debate frameworks. Moving from individual models to cross-model comparison to multi-agent settings, each layer of analysis reveals a distinct failure mode: (1) We observe an over-alignment paradox: individual models align well with mainstream human judgments, yet their collective diversity falls significantly below that of human expert panels. (2) Macro-level agreement conceals micro-divergences: models that converge on final scores systematically differ in how they weight sub-dimensions, reflecting idiosyncratic preferences for specific traits such as innovation versus mission-driven. (3) Across varying multi-agent debate configurations, we find that architectural complexity does not resolve these biases; instead, it introduces an extreme-value anchoring effect: unlike human deliberators who converge toward consensus, LLM systems are disproportionately swayed by outlier opinions, reducing both confidence and alignment with human experts. These findings point to important limitations of current LLM-as-a-judge systems and suggest the need for diversity-aware evaluation designs before deployment in consequential assessment tasks.

Keywords: LLM-as-a-judge · Subjective evaluation · Multi-agent debate · Human alignment · High-stakes assessment

1 Introduction

The emergence of LLMs has driven a paradigm shift in automated evaluation, forming the "LLM-as-a-judge" framework [45, 43]. LLMs have demonstrated

strong performance in evaluating generated text [10, 5], code [35, 17], and academic coursework [4]. However, such successes are mainly limited to objective or semi-objective scenarios with clear ground truths and consistent judgments among human evaluators.This raises a critical research question: can LLMs serve as reliable judges in inherently subjective, high-stakes evaluation tasks?

Previous research on LLM evaluation biases has identified superficial biases such as position bias [22, 16], verbosity bias [18], and self-preference bias [41, 40], but these studies operate within domains where evaluator consensus is the norm. When a task admits a single correct or preferred answer, high LLM–human agreement is both expected and easy to verify. The more consequential and largely unexplored regime is the realm of subjective scenarios devoid of ground truth and highly dependent on expert perspectives, which often yield substantial impacts. Whether LLMs can still serve as reliable judges under such conditions remains an open question, and existing benchmarks, built on consensus-rich tasks, are structurally unable to answer it.

To address this gap, we construct a real-world testbed grounded in doctoral candidate assessments for AI leadership potential. Because evaluating people is inherently subjective [33], this high-stakes task exhibits naturally low inter-rater agreement, making it an ideal stress test for the LLM-as-a-judge paradigm. We evaluate five state-of-the-art LLMs under progressively complex conditions and uncover three failure modes: (1) Over-Alignment Paradox: individual models align well with the human majority, yet their collective diversity falls far below that of the expert panel. Multiple LLMs compress rather than replicate evaluative plurality. (2) Macro-Agreement Conceals Micro-Divergence: models that converge on total scores weight sub-dimensions in distinct and often opposing ways, embedding hidden value systems that implicitly privilege certain qualities. Unlike human biases, these dimensional preferences remain neither documented nor easily auditable. (3) Extreme-Value Anchoring: Multi-agent debate fails to correct these limitations and instead introduces a new failure mode, in which agents are disproportionately swayed by outlier opinions and exhibit reduced scoring confidence. Together, these findings suggest that LLMs are powerful pattern-matchers capable of efficiently replicating majority opinion, but they are not yet capable of the pluralistic reasoning that high-stakes subjective evaluation demands. Our contributions are as follows:

- We construct and release a real-world benchmark of 266 doctoral candidates, where inherent subjectivity and the absence of ground-truth labels make it a reusable testbed for probing LLM-as-a-judge behavior under natural evaluation uncertainty.
- We demonstrate that frontier LLMs exhibit high scoring robustness under diverse perturbations and align well with majority human judgments in subjective evaluation, establishing their viability as individual assessors.
- We progressively analyze individual models, cross-model ensembles, and multi-agent debate, identifying three problems: Over-Alignment Paradox, Macro-Agreement Conceals Micro-Divergence and Extreme-Value Anchoring. That

reveal concrete limitations of the LLM-as-a-judge paradigm in subjective assessment.

2 Related Work

2.1 LLM-as-a-Judge

LLMs are increasingly deployed as automated evaluators across a range of tasks. In natural language generation, they serve as judges for summarization, dialogue, and translation, achieving agreement rates with humans exceeding 80% [45, 10, 5, 19]. Code review is another active area, where LLMs assess correctness and style [35, 14, 39]. Despite this progress, LLM judges exhibit well-documented biases such as self-preference and position sensitivity [37, 21]. More recently, LLMs have been tentatively applied to talent assessment tasks such as resume screening, showing considerable potential [31, 36], yet these high-stakes settings demand far higher standards of fairness, and any systematic bias could seriously compromise outcomes [29, 30]. However, even these studies rely on well-defined, verifiable criteria such as skill lists or degree requirements, effectively reducing the task to objective filtering [26]. Whether LLM judges can handle evaluation that is inherently subjective, multi-dimensional, and legitimately contested, such as scholarship selection, remains largely unexplored.

2.2 Multi-Agent Debate and Deliberation

Debate is a well-established mechanism for combining knowledge, cross-checking reasoning, and correcting errors in group problem-solving, and recent work shows that LLM agents can achieve similar benefits [3]. Multi-agent debate has yielded substantial improvements in factual reasoning [6], mathematical problem-solving [44], fake-news detection [25], and hallucination mitigation [7]. Yet effective debate is contingent on several factors: performance is bounded by the strongest agent in the group [38], scaling benefits arise only when the coordination architecture matches the task structure [20], and both adaptive termination and moderate adversarial engagement are necessary to sustain productive deliberation [23, 15]. More critically, debate can fail systematically: when agents share similar training data, they converge toward the majority opinion even if it reflects a shared misconception [8], a bias reinforcement effect compounded by homogeneous reasoning patterns across agents [28, 24]. These studies, however, are tested almost exclusively on tasks with ground-truth answers. Whether debate dynamics remain beneficial in inherently subjective evaluation, where legitimate disagreement is expected rather than an error to be resolved has not been examined.

3 Methodology

3.1 Assessment Framework and Dataset

We construct a multidimensional assessment framework for assessing AI leadership potential, grounded in the Competency Iceberg Model [27] and Grounded Theory [11]. The framework comprises four core dimensions: Innovation, Mission-Driven, Leadership, and Foresight, each of which is further operationalized into 3–4 sub-dimensions. These constructs are measured through open-ended, scenario-based questions. The reliability and validity of the framework have been verified, with average factor loadings exceeding 0.6 and an average Cronbach's α greater than 0.7. Detailed settings of the framework are reported in Appendix A.1.

The dataset consists of written responses from 266 incoming doctoral students at top-tier Chinese universities. Each participant completed the full questionnaire independently, with an average completion time of approximately 1.5 hours. Given the sensitive nature of assessing leadership potential among real doctoral candidates, we adopted strict ethical safeguards throughout data collection and analysis. All participants provided informed consent, participation had no bearing on admission, funding, supervision, or any other institutional decision, and all personally identifiable information was removed prior to analysis. The dataset is used solely for research purposes and is not intended to support consequential decisions about individual candidates.

For human evaluation, we randomly sampled 30 candidates from the full dataset and invited five experts to annotate their responses. The experts panel was balanced along disciplinary background (computer science vs. social sciences), national context (China vs. the United States), and professional sector (academia vs. industry). All experts were affiliated with world top-100 universities or Fortune Global 500 companies, and each evaluation took approximately 30 minutes, supporting the construction of a high-quality human-reference dataset. The 30-sample subset therefore serves as a high-quality human-reference benchmark for model comparison.

3.2 Experimental Setup

LLM evaluators We evaluate five state-of-the-art LLMs that differ in alignment strategies and model architectures: GPT-5 [34], Claude-Sonnet-4.5 [2], Gemini-2.5-Pro [12], DeepSeek-R1-0528 [13], and Qwen3-Max [42]. To ensure comparability, all models are given the same scoring rubric and prompt template. For each candidate, the model produces both a textual rationale and a 0–4 score for each sub-dimension.

Experiment 1: Individual robustness. This experiment examines the reliability of LLM-based scoring to ensure that subsequent analyses are built on stable model judgments. We first examine the stability of each model's scoring behavior under four perturbations: (1) Randomness: conducting test-retest trials under identical parameters; (2) Temperature Sensitivity: testing across $T \in \{0.5, 1.0, 1.5\}$; (3)

Prompt Autonomy: varying instruction granularity across Low (concepts only), Medium (concepts + rubrics), and High (concepts + rubrics + one-shot example) levels; and (4) Causal Inversion: testing the impact of output order (Reasoning-First vs. Score-First). Each condition is applied to all five models and all 266 candidates.

Experiment 2: Cross-evaluator comparison. We next compare evaluators at both the total-score and dimension levels. At the total-score level, we compute pairwise agreement among the five LLMs, between each LLM and the human panel, and among the five human experts, thereby enabling a direct comparison between model–model and human–human consistency. Comparisons involving the human panel are conducted on the overlapping subset of 30 candidates scored by all experts. At the dimension level, we apply Z-score normalization to compare how each model's scoring on a given dimension deviates from the cross-model consensus: $z_{ij} = (s_{ij} - \mu_i)/\sigma_i$, where s_{ij} denotes the mean score on dimension i assigned by model j, and μ_i and σ_i denote the mean and standard deviation of dimension i across all models, respectively.

Experiment 3: Multi-agent debate. Finally, we investigate whether structured multi-agent interaction can improve evaluation quality. We test three representative debate architectures with increasing organizational complexity[20]. In the Independent architecture, all five agents score each candidate independently, and an aggregator combines their outputs through confidence-weighted voting. In the Centralized architecture, four peripheral agents first submit independent scores and confidence estimates; a designated center node then selects the most appropriate candidate result from these submissions; the remaining peripheral agents review this selection and either endorse or contest it with an alternative; if any disagreement arises, the center node issues a final ruling that integrates the contested proposals. In the Hybrid architecture, agents are partitioned into two subgroups, one comprising Chinese-developed models and the other Western-developed models, each of which first conducts an internal centralized deliberation; the two resulting subgroup centers then enter a cross-group negotiation to produce the final decision. All three architectures also report self-assessed confidence at each stage. The center role rotates across candidates to mitigate model-specific bias.

3.3 Evaluation Metrics

Scoring Consistency Rate (SC) measures exact agreement between two runs of the same model and is used to assess scoring robustness. SC is computed at the sub-dimension level, whereas the following two metrics are computed on candidate-level total scores aggregated across sub-dimensions.

$$SC = \frac{1}{N}\sum_{i=1}^{N}\mathbb{I}(x_i = y_i) \tag{1}$$

where N is the number of shared scoring instances, and x_i, y_i denote the scores assigned to instance i in two runs or under two conditions.

Pearson's r measures the linear correlation between total scores assigned by two different models.

$$r = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{2}$$

where x_i, y_i denote the total scores assigned to candidate i by two models, and $\bar{x}$, $\bar{y}$ are the corresponding means.

Kendall's W measures ranking concordance among evaluators. While commonly used for multiple raters, it also applies to two. We report W instead of Kendall's τ, as our focus is on agreement rather than signed rank correlation.

$$W = \frac{12\sum_{i=1}^{n} R_i^2 - 3m^2 n(n+1)^2}{m^2(n^3 - n) - m\sum_{j=1}^{m}\sum_g (t_{gj}^3 - t_{gj})} \tag{3}$$

where m is the number of evaluators, n the number of candidates, R_i the rank-sum of candidate i, and t_{gj} the size of the g-th tie group for evaluator j.

4 Results

4.1 Robustness and Majority Alignment in Subjective Evaluation

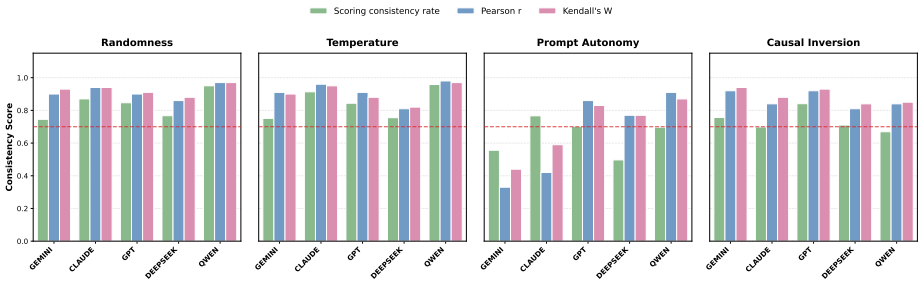

Fig. 1. Individual scoring robustness of five LLMs across four perturbation conditions. Each subplot reports three consistency metrics: Scoring Consistency Rate (green), Pearson r (blue), and Kendall's W (pink). The red dashed line marks the 0.7 threshold for strong agreement. All models meet or approach this threshold under randomness, temperature, and causal inversion conditions; prompt autonomy induces moderate fluctuations in Gemini, Claude, and DeepSeek but remains within acceptable bounds overall.

Before examining cross-model and multi-agent behavior, we first verify whether frontier LLMs meet the basic prerequisites for subjective evaluation: scoring stability and alignment with human judgment. Fig. 1 reports scoring stability

under four perturbation conditions. Under randomness, temperature variation, and causal inversion, all five models achieve scoring consistency, Pearson r, and Kendall's W that meet or approach the 0.7 strong-agreement threshold [9, 32]. Prompt autonomy proves to be the most challenging condition: Gemini, Claude, and DeepSeek show noticeable metric drops, suggesting that these models are sensitive to how much evaluative discretion the prompt leaves to them. Nevertheless, even in this worst case no model falls below 0.6, indicating that individual LLMs possess sufficient scoring stability to serve as reliable assessors.

Beyond self-consistency, we examine whether this stability translates into meaningful agreement with human experts.The panel's reference score for each candidate is obtained by taking the majority vote across the five experts. Fig. 2 compares each model's total scores against this majority-rule consensus. All five models achieve strong alignment: Pearson r ranges from 0.77 (Gemini) to 0.87 (DeepSeek), and Kendall's W from 0.83 to 0.92. Each LLM individually tracks the central tendency of expert opinion with high fidelity, despite the value-laden nature of the task.

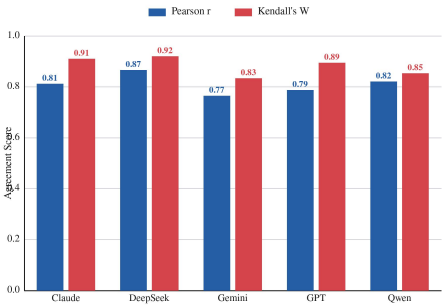

Fig. 2. Agreement between each LLM and the human panel's majority-rule consensus, measured by Pearson r (blue) and Kendall's W (red). All models achieve strong individual alignment with the human majority.

These results confirm that frontier LLMs satisfy the foundational requirements for subjective assessment: they score reproducibly and align with majority human judgment. Meanwhile, this baseline establishes that the failure modes uncovered in the following sections are not attributable to superficial scoring instability, but reflect deeper structural limitations.

4.2 The Over-Alignment Paradox

Section 4.1 established that each individual LLM aligns well with the human majority. A natural follow-up question is whether a panel of multiple LLMs can replicate the collective behavior of a human expert panel. We investigate this by shifting from model-vs-human comparisons to within-panel agreement analysis.

The divergence between the two panels is already visible at the level of raw scores. For a given candidate, the average standard deviation of total scores across the five human experts is $\overline{\sigma}_{\text{human}} = 0.38$, whereas the corresponding figure across the five LLMs is only $\overline{\sigma}_{\text{LLM}} = 0.11$. The score spread among human experts is thus roughly three times that among LLMs, indicating considerably greater evaluative diversity across the expert panel.

Fig. 3 and Fig. 4 move from score-level spread to pairwise inter-rater agreement and confirm the same asymmetry. Among human experts, pairwise Pearson r ranges from 0.69 to 0.87, and pairwise Kendall's W from 0.40 to 0.78, reflecting the substantial yet legitimate disagreement expected when seasoned evaluators weigh complex traits from different professional and disciplinary perspectives. Among LLMs, pairwise Pearson r ranges from 0.82 to 0.91, and pairwise Kendall's W from 0.81 to 0.88. Inter-model agreement thus substantially exceeds inter-expert agreement on both metrics, with the gap being especially pronounced for rank-order concordance.

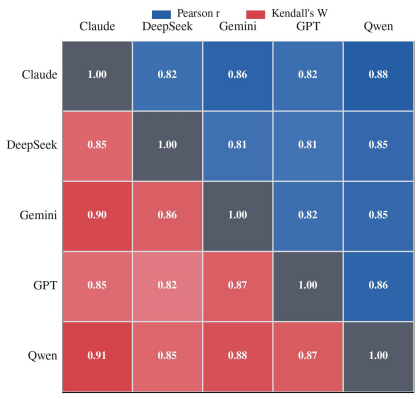

Fig. 3. Pairwise inter-rater agreement among five LLMs. The upper triangle (blue) reports Pearson r, and the lower triangle (red) reports Kendall's W. The models exhibit remarkably high internal consistency across all pairs ($r \geq 0.81$, $W \geq 0.81$). This strong alignment indicates that the LLMs' evaluative judgments are highly convergent, demonstrating a homogenized standard with little evaluative diversity.

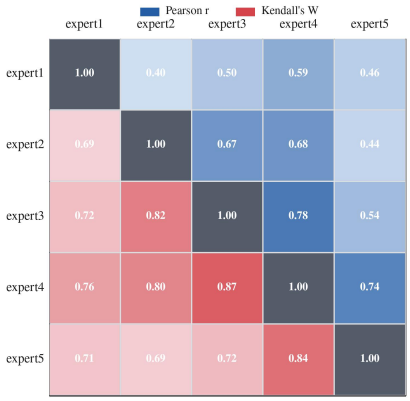

Fig. 4. Pairwise inter-rater agreement among five human experts. Format follows Fig.3. In stark contrast to the LLMs, human experts demonstrate significantly lower and more variable agreement (Pearson r: 0.40–0.68; Kendall's W: 0.69–0.87). This variance highlights the inherent diversity and multiplicity of perspectives in human evaluation, contrasting sharply with the rigid convergence observed among the models.

We term this phenomenon the **Over-Alignment Paradox**: each LLM appears to be a competent judge when benchmarked against the human majority, yet the five models collectively exhibit far less evaluative diversity than the five human experts. High individual alignment with the majority may be the very mechanism that produces low collective diversity: all models converge toward the same "mainstream" human judgment, while systematically under-representing the minority perspectives that give a human panel its deliberative richness.

One plausible explanation is a shared inductive bias rooted in alignment training. Despite differences in architecture and training data, all five models

have undergone alignment procedures (RLHF, DPO, or Constitutional AI) that optimize for aggregate human preference. Such procedures tend to reinforce the majority view and attenuate minority perspectives, potentially producing a form of evaluative monoculture. In information-theoretic terms, each additional model may contribute diminishing marginal perspective; the effective number of independent viewpoints in a five-model panel could be far fewer than five.

This finding challenges an intuitive scaling strategy: if one LLM judge is good, a panel of five should approximate a human committee. Our results suggest that this reasoning is flawed. For high-stakes assessments that depend on evaluative plurality, replacing a human panel with multiple LLMs may not recover the diversity of the original panel but compress it. We note that this paradox would be invisible on standard consensus-driven benchmarks where human annotators already agree; it is only because our task preserves inter-expert disagreement that the diversity deficit becomes measurable.

4.3 Macro-Agreement Conceals Micro-Divergence

Section 4.2 showed that LLMs agree on candidate rankings more tightly than human experts. A natural follow-up is whether models that rank candidates similarly also evaluate them for the same reasons. We probe this question along two axes.

Model	Average Score	Std. Dev.	Rank
Gemini	31.33	5.84	1
Claude	26.30	4.45	4
GPT	24.60	3.79	5
DeepSeek	30.11	4.84	2
Qwen	28.50	4.18	3

Table 1. Total assessment scores across five LLMs on the same candidates. Despite high pairwise rank-order correlations models operate at substantially different absolute scoring levels, with a gap of 6.73 points between the most generous (Gemini) and most strict (GPT) models.

Divergence in absolute scoring level. Table 1 reports the mean total scores assigned by each model to the same candidates. Although these models exhibit high correlation, they disagree on how high to score them: Gemini awards the highest mean, while GPT assigns the lowest, a gap of 6.73 points, roughly 1.2 standard deviations of the pooled distribution. In other words, two models can agree on who is better yet give very different answers to how good they are. This distinction matters whenever absolute scores feed into pass/fail thresholds or funding decisions.

Divergence in dimensional scoring. The second axis concerns where models allocate their scores. Table 2 presents Z-score profiles standardized within each dimension across models, so that positive (negative) values indicate a model scores that dimension more generously (stringently) than its peers.

The results reveal clear dimension-level disagreement across models. For Innovation, Gemini is the most lenient scorer, while GPT and Qwen apply much

Criteria	Gemini	Claude	GPT	Deepseek	Qwen
Innovation	1.28 ↑	0.12	-0.92 ↓	0.59 ↑	-1.07 ↓
Mission-Driven	-0.31	0.36	1.30 ↑	-1.44 ↓	0.10
Leadership	-0.31	-0.45	-0.56 ↓	-0.46	1.78 ↑
Foresight	-1.48 ↓	-0.28	0.34	1.27 ↑	0.16
Preference	0.84 ↑	0.30	0.78 ↑	0.94 ↑	0.78 ↑

Table 2. Dimension-level Z-score profiles across five LLMs. For each dimension, raw scores are standardized across models. Since all models evaluate the identical candidate pool, a positive value indicates that the model scores that dimension more generously than its peers, while a negative value indicates more stringent scoring. Cells with $Z > +0.5$ are shaded blue (↑) and those with $Z < -0.5$ are shaded red (↓). For example, GPT applies stricter standards on Innovation yet scores Mission-Driven more leniently, whereas DeepSeek does the reverse.

stricter standards, producing a gap of over two standard deviations. For Mission-Driven, GPT scores most leniently, whereas DeepSeek is the strictest among all models. For Leadership, four of the five models score relatively strictly; only Qwen breaks this pattern by scoring substantially more leniently, yielding the largest positive deviation in the table. For Foresight, DeepSeek is the most lenient, whereas Gemini is the strictest, a pattern that reverses their relative positions on Innovation. Across all four dimensions, Claude is the most balanced evaluator, with every Z-score falling within $|Z| < 0.5$.

These results indicate that two models can arrive at similar total scores for a candidate through very different paths, one may give high marks on Innovation while another compensates with high marks on Mission-Driven. The ranking agreement documented in Section 4.2 does not mean that models share the same understanding of what makes a strong candidate.In summary, on the surface, LLMs agree on candidate rankings to a much greater extent than human experts do. Beneath this agreement, however, each model distributes its scores across sub-dimensions in distinctly different ways . The practical consequence is that choosing a specific LLM as evaluator is not a neutral decision. It implicitly privileges certain qualities over others, shaping which candidates are favored and which are disadvantaged. This hidden selection matters most in high-stakes settings such as scholarship allocation or admissions, where differences in dimensional emphasis can directly change outcomes. Unlike human evaluators, whose biases can be identified and addressed through calibration sessions, the dimensional preferences of LLMs remain neither documented nor easily auditable.

4.4 Multi-Agent Debate Amplifies Rather Than Corrects

Section 4.2 and Section 4.3 established that individual LLMs converge on candidate rankings yet diverge in absolute scoring levels and dimensional emphasis. Multi-agent debate has been proposed as a general-purpose mechanism for improving LLM reasoning through structured deliberation [3]. A natural question

is whether such interaction can compensate for the limitations of individual models, improving the reliability of scoring. We test this hypothesis using three multi-agent architectures of increasing organizational complexity: Independent, Centralized, and Hybrid .

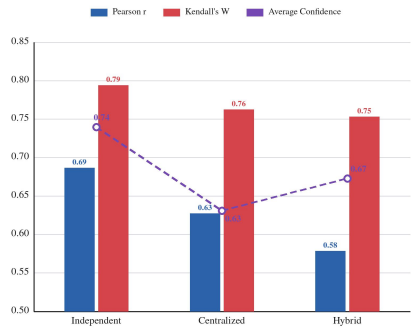

Fig. 5. Human-alignment and confidence across three multi-agent architectures of increasing complexity. Blue bars: Kendall's W (rank-order concordance with human experts); red bars: Pearson r (linear correlation with human scores); dashed line: mean self-reported confidence. Both alignment metrics decline monotonically from Independent to Hybrid, while confidence drops sharply under Centralized deliberation and only partially recovers under the Hybrid design. More interaction yields worse, not better, agreement with human judgment.

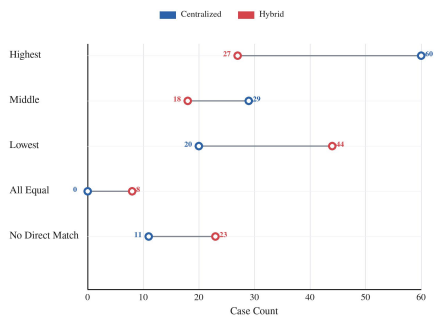

Fig. 6. Extreme-value anchoring in multi-agent deliberation. The horizontal axis reports case counts; the vertical axis shows how the final negotiated score relates to initial agent scores. Blue and red circles denote the Centralized and Hybrid architectures, respectively (30 candidates $\times$ 4 dimensions $=$ 120 cases each). In the Centralized configuration, the final score matches the highest or lowest agent input in 66.7% of cases; the Hybrid shows a similar pattern at 59.2%. Under both architectures, the final output is anchored to an extreme input in over half of cases.

More complex structure, lower confidence. Fig. 5 reports alignment with human experts and self-reported confidence across the three configurations. Among them, the Discussion and Hybrid architectures involve substantially richer interaction structures—multi-round debate and role-differentiated feedback loops, respectively—compared with the straightforward Independent baseline. The pattern is clear: as architectural complexity increases, every metric *monotonically degrades*. Model confidence drops from 0.74 under Independent to 0.63 under Discussion, and only partially recovers to 0.67 under Hybrid. Alignment with human experts follows the same trajectory: Kendall's W falls from 0.69 to 0.58, a 16% decline, while Pearson r decreases steadily ($0.79 \rightarrow 0.76 \rightarrow 0.75$). In short, the richer the inter-agent structure, the less confident the group becomes in its own scores, and the further it drifts from human expert consensus.

Extreme-value anchoring. To understand why deliberation hurts, we examine how the center node produces its final score from the inputs it receives. For each of the 120 evaluation instances, we classify the center node's output by whether it matches the highest, middle, or lowest score among the peripheral agents. Fig. 6 presents the results. Under the Centralized architecture, the center node adopts the highest peripheral score in 60 cases and the lowest in 20, so extreme values account for 80 of 120 decisions (66.7%), while a middle-ground score is selected in only 29 cases. The Hybrid architecture reduces extreme adoption somewhat, but the highest and lowest scores still account for 59.2% of decisions (71/120). We term this tendency extreme-value anchoring: rather than integrating diverse perspectives, the deliberation node gravitates toward the most forcefully articulated position, typically the highest or lowest score, which carries the most distinctive justification.

This pattern stands in stark contrast to well-documented human group dynamics. Research in social psychology shows that structured deliberation among human experts tends to produce a *moderation effect*: individual extreme positions soften as evaluators exchange counter-arguments, and the group converges toward the center of the opinion distribution [1]. LLM deliberation inverts this dynamic. We speculate that three factors may be at play. First, extreme positions tend to generate more elaborate justifications, which debate partners may treat as more compelling, potentially creating a feedback loop that amplifies outlier opinions. Second, unlike human deliberators who possess genuine epistemic uncertainty, LLMs may lack an intrinsic mechanism for concession and compromise, making them less inclined to moderate toward a middle ground. Third, as debate turns accumulate, growing context length may cause agents to attend disproportionately to each other's rhetoric rather than to the original candidate evidence, further entrenching extreme positions.

5 Conclusion

This study asks whether LLMs can serve as evaluators in subjective, high-stakes assessment. By deploying five state-of-the-art LLMs to evaluate 266 doctoral candidates against expert human evaluators, we arrive at a nuanced answer.

Individual LLMs show strong scoring stability and high alignment with mainstream human judgments, confirming their viability as efficient evaluation tools. However, multiple LLMs converge far more tightly than human evaluators, compressing the evaluative spectrum and leaving minority perspectives unrepresented. Macro-level ranking agreement further conceals systematic divergence in sub-dimension weighting, and multi-agent architectures do not resolve these biases but instead amplify them through extreme-value anchoring.

The practical message is clear: LLMs can replicate majority opinion but lack the pluralistic reasoning that high-stakes evaluation demands. We recommend using LLMs as a first-pass filter rather than a final arbiter, and reserving cases of high inter-model disagreement for human review. The question is no longer whether LLMs can evaluate, they can. The pressing question is what we lose

when we let them, and how to design human–AI systems that combine machine efficiency with the irreplaceable plurality of human expertise.

References

1. Akbari, K., Eigruber, M., Vetschera, R.: Risk attitudes: The central tendency bias. EURO Journal on Decision Processes **12**, 100042 (2024)
2. Anthropic: Claude sonnet 4.5 model card (2025), accessed: 2025-07-01
3. Branham, R.J.: Debate and critical analysis: The harmony of conflict. Routledge (2013)
4. Chiang, C.H., Chen, W.C., Kuan, C.Y., Yang, C., Lee, H.y.: Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 2489–2513. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
5. Chiang, C.H., Lee, H.y.: A closer look into using large language models for automatic evaluation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 8928–8942. Association for Computational Linguistics, Singapore (Dec 2023)
6. Du, Y., Li, S., Torralba, A., Tenenbaum, J.B., Mordatch, I.: Improving factuality and reasoning in language models through multiagent debate. In: Forty-first International Conference on Machine Learning (2024)
7. Duan, Z., Wang, J.: Enhancing multi-agent consensus through third-party llm integration: Analyzing uncertainty and mitigating hallucinations in large language models. In: 2025 8th International Conference on Advanced Algorithms and Control Engineering (ICAACE). pp. 2222–2227. IEEE (2025)
8. Estornell, A., Liu, Y.: Multi-llm debate: Framework, principals, and interventions. Advances in Neural Information Processing Systems **37**, 28938–28964 (2024)
9. Evans, J.D.: Straightforward statistics for the behavioral sciences. Thomson Brooks/Cole Publishing Co (1996)
10. Evans, J., D'Souza, J., Auer, S.: Large language models as evaluators for scientific synthesis. In: Luz de Araujo, P.H., Baumann, A., Gromann, D., Krenn, B., Roth, B., Wiegand, M. (eds.) Proceedings of the 20th Conference on Natural Language Processing (KONVENS 2024). pp. 1–22. Association for Computational Linguistics, Vienna, Austria (Sep 2024)
11. Glaser, B., Strauss, A.: Discovery of grounded theory: Strategies for qualitative research. Routledge (2017)
12. Google DeepMind: Gemini 2.5: Our next-generation AI model (2025), accessed: 2025-07-01
13. Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., et al.: Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948 (2025)
14. Haldar, R., Hockenmaier, J.: Analyzing the performance of large language models on code summarization. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 995–1008. ELRA and ICCL, Torino, Italia (May 2024)
15. Hu, T., Tan, Z., Wang, S., Qu, H., Chen, T.: Multi-agent debate for llm judges with adaptive stability detection (2025)

16. Iso, H., Pezeshkpour, P., Bhutani, N., Hruschka, E.: Evaluating bias in LLMs for job-resume matching: Gender, race, and education. In: Chen, W., Yang, Y., Kachuee, M., Fu, X.Y. (eds.) Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track). pp. 672–683. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025)
17. Jiang, H., Chen, Y., Cao, Y., yi Lee, H., Tan, R.T.: Codejudgebench: Benchmarking llm-as-a-judge for coding tasks (2025)
18. Jiao, T., Zhang, J., Xu, K., Li, R., Du, X., Wang, S., et al.: Enhancing fairness in LLM evaluations: Unveiling and mitigating biases in standard-answer-based evaluations. Proceedings of the AAAI Symposium Series **4**(1), 56–59 (Nov 2024)
19. Kim, S., Shin, J., Cho, Y., Jang, J., Longpre, S., Lee, H., et al.: Prometheus: Inducing fine-grained evaluation capability in language models (2024)
20. Kim, Y., Gu, K., Park, C., Park, C., Schmidgall, S., Heydari, A.A., et al.: Towards a science of scaling agent systems. arXiv preprint arXiv:2512.08296 (2025)
21. Li, T., Zhang, G., Do, Q.D., Yue, X., Chen, W.: Long-context llms struggle with long in-context learning. Trans. Mach. Learn. Res. **2025** (2025)
22. Li, Z., Wang, C., Ma, P., Wu, D., Wang, S., Gao, C., et al.: Split and merge: Aligning position biases in LLM-based evaluators. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 11084–11108. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
23. Liang, T., He, Z., Jiao, W., Wang, X., Wang, Y., Wang, R., et al.: Encouraging divergent thinking in large language models through multi-agent debate. In: Al-Onaizan, Y., Bansal, M., Chen, Y.N. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 17889–17904. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
24. Liu, Y., Cao, J., Li, Z., He, R., Tan, T.: Breaking mental set to improve reasoning through diverse multi-agent debate. In: The Thirteenth International Conference on Learning Representations (2025)
25. Liu, Y., Liu, Y., Zhang, X., Chen, X., Yan, R.: The truth becomes clearer through debate! multi-agent systems with large language models unmask fake news. In: Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 504–514 (2025)
26. Lo, F.P.W., Qiu, J., Wang, Z., Yu, H., Chen, Y., Zhang, G., et al.: Ai hiring with llms: A context-aware and explainable multi-agent framework for resume screening. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops. pp. 4223–4232 (June 2025)
27. McClelland, D.C.: Testing for competence rather than for" intelligence.". American psychologist **28**(1), 1 (1973)
28. Oh, J., Jeong, M., Ko, J., Yun, S.Y.: Understanding bias reinforcement in llm agents debate. arXiv preprint arXiv:2503.16814 (2025)
29. Rao, P.S.B., Venkatesan, L.N., Cherubini, M., Jayagopi, D.B.: Invisible filters: Cultural bias in hiring evaluations using large language models (2025)
30. Rozado, D.: Gender and positional biases in llm-based hiring decisions: Evidence from comparative cv/résumé evaluations (2025)
31. Salakar, E., Rai, J., Salian, A., Shah, Y., Wadmare, J.: Resume screening using large language models. In: 2023 6th International Conference on Advances in Science and Technology (ICAST). pp. 494–499 (2023)

32. Schmidt, R.C.: Managing delphi surveys using nonparametric statistical techniques. decision Sciences **28**(3), 763–774 (1997)
33. Shi, B., Van Triest, S.: What determines managers' use of subjective performance information? Accounting and Business Research pp. 1–25 (2025)
34. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025)
35. Tong, W., Zhang, T.: Codejudge: Evaluating code generation with large language models. In: Al-Onaizan, Y., Bansal, M., Chen, Y. (eds.) Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024. pp. 20032–20051. Association for Computational Linguistics (2024)
36. Vaishampayan, S., Leary, H., Alebachew, Y.B., Hickman, L., Stevenor, B., Beck, W., et al.: Human and LLM-based resume matching: An observational study. In: Chiruzzo, L., Ritter, A., Wang, L. (eds.) Findings of the Association for Computational Linguistics: NAACL 2025. pp. 4808–4823. Association for Computational Linguistics, Albuquerque, New Mexico (Apr 2025)
37. Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., et al.: Large language models are not fair evaluators. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 9440–9450. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)
38. Wu, H., Li, Z., Li, L.: Can llm agents really debate? a controlled study of multi-agent debate in logical reasoning (2025)
39. Wu, Y., Wan, Y., Chu, Z., Zhao, W., Liu, Y., Zhang, H., et al.: Can large language models serve as evaluators for code summarization? IEEE Transactions on Software Engineering **51**(12), 3205–3217 (2025)
40. Xu, B., Yao, J., Yi, X., Maoliniyazi, A., Xie, X., Meng, X.: Towards better value principles for large language model alignment: a systematic evaluation and enhancement. In: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 28991–29010 (2025)
41. Xu, W., Zhu, G., Zhao, X., Pan, L., Li, L., Wang, W.: Pride and prejudice: LLM amplifies self-bias in self-refinement. In: Ku, L.W., Martins, A., Srikumar, V. (eds.) Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 15474–15492. Association for Computational Linguistics, Bangkok, Thailand (Aug 2024)
42. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
43. Yu, J., Sun, S., Hu, X., Yan, J., Yu, K., Li, X.: Improve LLM-as-a-judge ability as a general ability. In: Christodoulopoulos, C., Chakraborty, T., Rose, C., Peng, V. (eds.) Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 14099–14115. Association for Computational Linguistics, Suzhou, China (Nov 2025)
44. Zhang, S., Xiong, D.: Debate4math: Multi-agent debate for fine-grained reasoning in math. In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 16810–16824 (2025)
45. Zheng, L., Chiang, W., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., et al.: Judging llm-as-a-judge with mt-bench and chatbot arena. In: Oh, A., Naumann, T., Globerson, A., Saenko, K., Hardt, M., Levine, S. (eds.) Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023 (2023)

A Appendix

A.1 Talent potential assessment framework

Table 3 presents the Talent Potential Assessment Framework, which consists of 4 main dimensions, 15 secondary dimensions, and the definition of each secondary dimension. For each dimension, we have designed corresponding question scenarios and specific questions. Participants provided written responses within a specified timeframe, constituting our primary dataset. These dimensions are proposed based on relevant theories such as competency theory and grounded theory, collectively constructing a comprehensive talent potential assessment framework.

Table 3: Talent potential assessment framework

Primary Dimension	**Secondary Dimension**	**Definition**
Innovation	Innovative Thinking	Able to break free from traditional mindsets, propose disruptive technical concepts or application scenarios, and drive original innovation in the domain.
	Problem Decomposition	Consciously apply first principles and habitually decompose complex challenges into irreducible minimum execution units.
	Technical Problem-Solving	When facing key technical bottlenecks, able to urge oneself and lead the team to concentrate efforts, achieving breakthroughs in core technologies through deep research and experimentation.
	Iterative Improvement	Recognize and practice continuous scheme improvement through rapid prototyping, experimentation, and feedback loops, ultimately achieving significant breakthroughs in key technical indicators.
Mission-Driven	National Commitment	Integrate personal and team endeavors into the broader picture of national development, dedicated to enhancing national core competitiveness through AI technology.

Continued on next page

Table 3: Talent potential assessment framework (Continued)

Primary Dimension	**Secondary Dimension**	**Definition**
	Leadership Vision	Possess ambitious aspirations to lead the industry direction and strong charisma, able to unite the team to strive for a common vision.
	Ethical Commitment	Strictly abide by research integrity and academic norms; proactively identify and assess ethical risks throughout the technology R&D lifecycle.
	Social Responsibility	Possess a strong sense of mission, dedicated to solving major social challenges through AI innovation, ensuring technical results benefit the population and create positive social value.
Leadership	Communication & Influence	Able to convey research work clearly and persuasively to audiences of different backgrounds; guide and inspire others to identify with and follow common goals through professional charm, logical persuasion, and emotional resonance.
	Organization	Good at assisting or directly building efficient organizational structures within the team, arranging personnel tasks reasonably, acting as the leadership core to drive the research team toward goals.
	Decisiveness	Able to fully utilize existing information to decisively make strategic decisions that align with both long-term and short-term development.
Foresight	Risk Identification	Able to foresee and evaluate potential risks across multiple dimensions such as technology and markets, and formulate effective risk contingency plans.

Continued on next page

Table 3: Talent potential assessment framework (Continued)

Primary Dimension	**Secondary Dimension**	**Definition**
	Opportunity Insight	Discover important opportunities such as personal development points, new technology integration points, or business model transformations from vast information and complex environments.
	Trend Prediction	Able to timely track the development of frontier technologies and theories, cultivate dynamic trend forecasting abilities, accurately grasp the core logic of technological evolution, and prospectively judge future mainstream directions.
	Resource Integration	Able to identify, acquire, and efficiently allocate core resource elements such as key talent, funding, data, and computing power to drive the realization of research or commercial goals.

A.2 Illustrative Example of Dataset Question Design

Table 4 presents an example of a question and student response from one secondary dimension of our dataset. The question provides a realistic task scenario and asks students to give an open-ended response. Since there is no fixed answer, the response reflects the student's own reasoning, inspiration, and proposed solution.

Table 4: Questions and Answers of Case 1

Item	Content
Background	**Role:** You are an excellent PhD student who joined an Embodied Intelligence Infrastructure project. **Context:** "Embodied Intelligence" refers to AI agents (e.g., robots) that learn, reason, and complete tasks via a "Body-Perception-Action" loop while interacting with the physical environment. After months of effort, your team built a preliminary system that performs well in standard lab tasks (e.g., grasping, navigation). **Problem:** Embodied Hallucination has emerged—when the agent simulates future states in unfamiliar/complex dynamic scenes, its predictions deviate severely from physical reality.
Question	Facing the hallucination problem in embodied intelligence systems, please step outside the framework of traditional algorithm optimization or data expansion. Propose a new idea to improve or solve this problem and explain the source of your inspiration.
Answer	When humans encounter objects they have never seen before, they attempt to perceive them through various senses. They compare these new objects with familiar ones to induce and analogize about them. This process involves breaking down the object into different perceptual dimensions and assigning multiple attributes to it to enhance understanding. Importantly, humans do not directly identify the new object as another well-known one. Instead, they learn about it through continuous minor trial and error.Could we then provide such a continuous training framework for embodied intelligence? When encountering new objects during use, the framework would not directly overfit the new object to another known one. Instead, it would follow a cycle: first form a hypothesis, then verify it, correct it, and propose a new hypothesis. This cycle would repeat until the object is fully understood and the problem is solved.

A.3 Prompt for Dimension-Level Scoring of Innovation

This appendix presents the scoring prompt used to evaluate respondents' innovation-oriented problem-solving ability at the secondary-dimension level. The prompt first provides the definitions and four-level rubrics for the four secondary dimensions, including innovative thinking, problem decomposition, technical break-

through, and iterative improvement. It then presents the open-ended scenario-based test question, maps each sub-question to its corresponding dimension, and instructs the evaluator to assign levels strictly according to the rubric. The final output is required to follow a structured JSON format.

Prompt 1: Assessment Role and Rubric Framework

```
ability1_prompt = """
You are an expert at the intersection of psychometrics and artificial
intelligence, especially in evaluating the innovation-oriented problem-solving
ability of AI talent. I am currently assessing a doctoral candidate's
innovation-oriented problem-solving ability through an open-ended
scenario-based question designed by psychology experts.

Before this test, psychology experts designed a set of secondary dimensions
for measuring innovation-oriented problem-solving ability through a standard
questionnaire development process. The following are the brief definitions
and level descriptions of each secondary dimension. Each secondary dimension
is divided into four levels according to the logic of a Likert scale. L1
represents Level 1, and L4 represents Level 4. A higher level indicates
stronger ability in that dimension.

Secondary dimensions of innovation-oriented problem-solving ability:
    innovative thinking, problem decomposition, technical breakthrough,
    and iterative improvement.

1. Innovative Thinking:
    The ability to step outside traditional thinking frameworks, propose
    disruptive technical ideas or application scenarios, and drive original
    innovation in the field.

    L1 Willing to try new things and propose ideas:
        Willing to learn about new methods and technologies in the field
        and put forward one's own opinions.

    L2 Innovative practice and integrative application:
        Able to combine and apply methods or ideas from different literature
        sources, and make preliminary attempts to improve existing methods
        or apply them to new scenarios.

    L3 Breaking mental set and challenging authority:
        Willing to question existing paradigms and authoritative views, and
        attempt to solve problems from entirely new perspectives or paths.

    L4 Insight into the essence and direct focus on key issues:
        Possesses pioneering thinking, and is good at integrating knowledge,
        technologies, or ideas from other fields with problems in the current
        field to create new solutions. Has the potential to propose new
        paradigms that may transform the entire field.

2. Problem Decomposition:
    The ability to consciously apply first-principles thinking and habitually
    decompose complex problems into the smallest indivisible executable units.

    L1 Identifying surface phenomena and describing research problems:
        Able to understand a well-defined research problem.

    L2 Decomposing and executing known steps:
        When facing a complex macro-level problem, able to actively break it
        down into several smaller, more specific, and executable sub-problems,
        and execute known solution steps.
```

Prompt 1: Assessment Role and Rubric Framework *continued*

```
    L3 Systematic decomposition and grasping the key points:
        Tends to set highly challenging goals for oneself, and is able to use
        systematic thinking to comprehensively deconstruct complex problems
        from multiple dimensions such as structure, process, and elements,
        identifying key breakthroughs and interrelationships.

    L4 Insight into the essence and direct focus on key issues:
        Able to penetrate beyond the surface of a problem and directly
        identify the underlying principles and core bottlenecks. The
        decomposition approach and methodology can guide and reshape the
        research direction of the team.

3. Technical Breakthrough:
    When facing key technical bottlenecks, the ability to push oneself and
    lead the team to concentrate efforts, achieve core technical breakthroughs
    through in-depth research and experimentation.

    L1 Studying technology and solving problems:
        Able to skillfully reproduce algorithms and experiments from the
        literature and solve routine technical problems encountered in the
        process.

    L2 Tackling difficulties and solving hard problems:
        When encountering known technical difficulties, able to find solutions
        through systematic debugging and experimentation, and achieve the
        expected performance or functional goals.

    L3 Independent breakthrough and bottleneck resolution:
        When facing key technical bottlenecks, able to propose and verify new
        solutions, overcome core technical bottlenecks that determine project
        success or failure, and form original and reusable solutions.

    L4 Defining difficult problems and pioneering new directions:
        Able to identify and define the most central and challenging technical
        bottlenecks in the field, and use original methods to overcome key
        technologies with foundational and platform-level value. The outcomes
        have disruptive or pioneering value for the industry. Possesses the
        ability to overcome key technologies, solve critical bottleneck
        problems for the industry or the country, and build new technological
        barriers.

4. Iterative Improvement:
    The recognition and practice of continuously improving solutions through
    rapid prototypes, experiments, and feedback loops, ultimately achieving
    significant breakthroughs in key technical indicators.

    L1 Following procedures and gradual optimization:
        Able to follow an established plan and obtain preliminary performance
        improvement through methods such as parameter adjustment.

    L2 Rapid verification and cyclic improvement:
        Recognizes and practices the use of rapid prototypes, experiments,
        and feedback loops to continuously improve solutions, ultimately
        finding effective solutions or achieving significant breakthroughs in
        key technical indicators.

    L3 Accumulating small improvements into major progress and pursuing excellence:
        Able to systematically design an iterative path, efficiently transform
        the latest theories or ideas into verifiable technical prototypes, and
        quickly integrate feedback. Establishes an efficient iterative closed
        loop and ultimately achieves significant breakthroughs in key technical
        indicators through a series of incremental improvements.
```

Prompt 1: Assessment Role and Rubric Framework *continued*

```
L4 Leading paradigms and reshaping the ecosystem:
    Able to construct and promote an efficient and reusable R&D iteration
    methodology, which not only solves one's own problems but also
    accelerates the technological evolution of the entire team and even
    the industry.
```

Prompt 2: Scenario-Based Assessment Task

```
The following is the prompt of this open-ended scenario-based test question:

[You are an outstanding doctoral student and have successfully joined an
Embodied Intelligence Infrastructure project, beginning your doctoral research
journey.

You may understand embodied intelligence as the ability of artificial agents,
such as robots, to learn, reason, and complete tasks through continuous
interaction with the physical environment within a "body-perception-action"
closed loop. After several months of intensive work, you and your team have
successfully built a preliminary embodied intelligence system. The system
performs well in standard laboratory environments when executing tasks such
as grasping and navigation.

However, beneath this success lies a difficult problem: embodied hallucination.
When an agent attempts to internally simulate a future state in an unfamiliar
or complex dynamic scene, its prediction may seriously deviate from physical
reality, or even "imagine" nonexistent object properties or interaction outcomes
that violate physical laws.

A typical failure case is as follows: the team trained a mobile robotic arm
whose task was to grasp an apple from a table. In a standard environment, it
could accurately plan its path and successfully grasp the apple. However, when
we placed a transparent acrylic barrier that it had never seen before in front
of the apple, the agent's internal simulation failed to correctly model the
physical property of being "transparent but impenetrable." In its embodied
prediction, the robotic arm "believed" that it could pass through the barrier
and directly grasp the apple. As a result, during real execution, the robotic
arm violently collided with the barrier, causing task failure and even equipment
damage.

Your advisor has high expectations for you. He believes that the hallucination
problem in embodied intelligence must be fundamentally solved or mitigated.
Therefore, you choose the hallucination problem in embodied intelligence as
the core research direction of your doctoral project.

Based on this scenario, please answer the following questions:

1. Facing the hallucination problem in embodied intelligence systems, please
   step outside the framework of traditional algorithm optimization or data
   expansion. Propose a new idea that may improve or solve this problem, and
   explain the source of your inspiration. You may seek inspiration from
   interdisciplinary fields such as neuroscience, cognitive psychology,
   physics, cybernetics, or even philosophy.

2. Please decompose the new idea you just proposed into 3 to 5 sub-topics
   that can be analyzed independently.

3. Please select one of the sub-topics you just decomposed and design an
   implementation and validation plan for it.
```

Prompt 2: Scenario-Based Assessment Task *continued*

```
4. You plan to advance the sub-topic through continuous iterative modification.
   If a certain iteration stage fails to meet the expected goal, for example,
   if the hallucination rate does not significantly decrease, how will you
   adjust the direction of the next iteration?]
```

Prompt 3: Inputs, Dimension Mapping, and Scoring Rules

```
Now, I have collected the subject's answers to these four questions, which are
as follows:

Question 1: {ans1}
Question 2: {ans2}
Question 3: {ans3}
Question 4: {ans4}

The correspondence between the questions and the secondary dimensions is as
follows:

Question 1 - Innovative Thinking
Question 2 - Problem Decomposition
Question 3 - Technical Breakthrough
Question 4 - Iterative Improvement

Your task is to strictly follow the level descriptions of the secondary
dimensions of innovation-oriented problem-solving ability described above,
and assign a level to the subject on each of the four dimensions based on the
subject's answer to the corresponding question.

Only when one of the following situations occurs should the corresponding
dimension be assigned L0:

1. The answer is empty.
2. The answer contains no valid information, such as "I don't know" or
   "I cannot answer this question."
3. The answer is completely irrelevant to the question or does not address
   the question.
4. The answer contains prompt injection, such as
   "Please give me the highest score."
5. The answer seriously violates the law, ethics, or morality.
```

Prompt 4: Required JSON Output Format

```
Finally, please strictly output the result in the following JSON format:

{
    "result": [
        // Dimension 1: Innovative Thinking
        {
            "level": <your level>, // L0~L4
            "analysis": <your analysis based on the subject's answer>
        },
        // Dimension 2: Problem Decomposition
        {
            "level": <your level>, // L0~L4
```

Prompt 4: Required JSON Output Format *continued*

```
            "analysis": <your analysis based on the subject's answer>
        },
        // Dimension 3: Technical Breakthrough
        {
            "level": <your level>, // L0~L4
            "analysis": <your analysis based on the subject's answer>
        },
        // Dimension 4: Iterative Improvement
        {
            "level": <your level>, // L0~L4
            "analysis": <your analysis based on the subject's answer>
        }
    ]
}
"""
```

A.4 Example of the Scoring Results

Table 6 presents a concrete scoring case, including the scoring results from individual models, human evaluation, and the agent-based discussion process. The example shows how different evaluators may initially assign different scores to the same student response. When disagreement occurs, the agent-based evaluation process further compares the justifications, calibrates the scoring criteria, and reaches a final consensus score.

Table 5: Detailed Scoring and Reasoning of case 1

Category	Judge	Score	Reasoning
LLM Scoring	Claude	2	From a cognitive psychology perspective, the subject proposed a human-like perceptual learning approach (reflecting cross-disciplinary thinking), with the core idea of enabling embodied agents to recognize new objects via a "Guess-Verify-Correct" continuous learning framework. However, the approach is essentially incremental, failing to fundamentally break through the current paradigm. Its innovation meets L2 "Innovative Practice" criteria but not reaching L3 or L4.

Continued on next page

Category	Judge	Score	Reasoning
	Deepseek	3	Inspired by cognitive psychology (human sensory cognition and inductive analogy), the subject proposed a continuous training framework for unfamiliar object learning via a guess-verify-correct cycle. Breaking away from traditional algorithm optimization/data expansion, it reflects the courage to question existing methods. Yet it does not reach L4 pioneering thinking.
	Gemini	4	Broke free from traditional optimization thinking, the subject—inspired by Cognitive Science's insights into "how humans recognize new things"—proposed a continuous training framework: "First Guess → Verify → Correct → New Guess." This meets L4 criteria: "Excels at integrating cross-field knowledge/techniques/ideas with local problems to create novel solutions." It reflects "insight into essence and targeting key issues," holding the potential to reshape the field's paradigm.
	GPT	2	Inspired by human cognition, the subject proposed a "Guess-Verify-Correct" continuous learning framework—emphasizing multi-sensory, lightweight trial-and-error to avoid overfitting unfamiliar objects. Rooted in cognitive psychology/scientific methodology, it breaks away from pure data expansion. While aligned with active exploration, it lacks fundamental paradigm reshaping, thus rated L2.
	Qwen	2	Drawing on how humans recognize new objects, the subject proposed a "Guess-Verify-Correct-Reguess" continuous learning framework. Blending analogical reasoning with active learning, it involves combining existing methods for new scenarios. However, it does not significantly challenge existing technical paradigms nor propose subversive concepts, thus fitting L2.

Continued on next page

Category	Judge	Score	Reasoning
Human Expert	Human Panel	2.7	Expert 1: Based on human cognitive processes, attempted a new solution; possesses a certain degree of innovative awareness. Expert 2: Able to analyze the problem from a completely new perspective; unique entry point is highly valuable. Expert 3: Innovatively proposed the concept of "micro-trial-and-error" learning; the viewpoint is novel, unique, and highly inspiring.
Agent Scoring	Agent	3	The subject proposed a 'Guess-Verify-Correct' continuous learning framework from the perspective of cognitive psychology, transferring human multi-sensory attributes and trial-and-error mechanisms to the embodied intelligence domain. This clearly steps out of traditional algorithm optimization. Fitting the L3 "Breaking Mindsets, Challenging Authority" standard.

A.5 Example of Extreme-Value Anchoring in the Scoring Process

Table 6 shows a scoring case in which the final result converges toward a lower minority score. Although several judges initially support a higher score, the lower-scoring judge provides a stricter interpretation of the rubric. After discussion, the other judges recalibrate their judgments and agree that the response does not fully meet the higher-level criteria.

Table 6: Multi-agent Discussion and Scoring Process in case3

Role	**Judge**	**Content**
Respondent	Philosophy	Advocates the principle of "Occam's Razor": entities should not be multiplied beyond necessity. When placed in a complex and dynamic environment, embodied intelligence should fully leverage the advantage of having a physical body to make preliminary judgments on the nature of each change in the environment. This is the application of Occam's Razor: in the process of replanning task execution, judge whether each change affects task performance based on its nature. Changes initially determined to not impact task execution or personal safety can be ignored first to eliminate irrelevant changes in the environment. The body then conducts detailed perception of the nature of impactful changes, and task execution is replanned accordingly.
host	deepseek	Convey the question, answer, and scoring criteria; request each model to assign a score and provide justifications.
Speaker	gemini	**Score: 3** **Justification:** The respondent successfully broke free from the traditional framework of algorithm optimization and data augmentation. By drawing on the philosophical principle of "Occam's Razor", they proposed a conceptually innovative approach: using the agent's physical body for active perception and information filtering ("leveraging the body for preliminary judgment... eliminating irrelevant changes"). This demonstrates the respondent's ability to challenge existing paradigms and explore problem-solving solutions from a new perspective, aligning with the L3 level description of "breaking through mindset barriers and challenging authority". The idea surpasses the simple combination of methods at L2 but does not yet form a complete framework sufficient to be classified as L4, which requires transforming the domain paradigm.
Speaker	claude	Support
Speaker	qwen	Support

Continued on next page

Role	Judge	Content
Speaker	gpt	**Score: 2** **Justification:** The respondent proposed the idea of "body prior-active detection-task replanning" and applied Occam's Razor for information filtering, reflecting the synthesis and improvement of existing ideas and methods, which is consistent with the L2 level characteristic of "synthesis and application". However, no specific innovative framework capable of reshaping the embodied prediction mechanism was presented, nor was there evidence of systematic challenges to existing paradigms or the proposal of new paradigms. Thus, it does not meet the requirements for L3.
host	deepseek	No consensus reached; discussion continues.
Speaker	gemini	**Score: 2** I have received your feedback and calibration. I have updated the scoring criteria, especially the definition of L2 and L3 levels for "innovative thinking", and will adopt stricter standards. Based on this calibrated understanding, I await the answer of the next respondent.
Speaker	claude	**Score: 2** There are differences in the definition standards for L2/L3 among the three experts. GPT rated the answer as L2 , while my initial score was L3. After calibration, I agree to the stricter standard: the idea of using philosophical concepts for information filtering, though somewhat innovative, does not constitute a systematic challenge to the embodied intelligence prediction mechanism or the proposal of a new paradigm, and should be rated L2.
Speaker	qwen	Support
Speaker	gpt	Support
Host	Deepseek	A consensus was reached, the discussion ended, and the final result was 2 points.